Xi and Zhou *Big Data Analytics* (2017) 2:9
DOI 10.1186/s41044-017-0023-7

Big Data Analytics

RESEARCH Open Access

Building a Chinese discourse topic corpus with a micro-topic scheme based on theme-rheme theory

Xue-feng Xi[1,2,3] and Guodong Zhou[2*]

*Correspondence:
gdzhou@suda.edu.cn
[2]School of Computer Science and Technology, Soochow University, ShiZi Road, Suzhou, China
Full list of author information is available at the end of the article

Abstract

Background: How to build a suitable discourse topic structure is an important issue in discourse topic analysis, which is the core of natural language understanding. Not only is it the key basic unit to implement automatic computing, but also the key to realize the transformation from unstructured data to structured data during the process of big data analytics. Although the discourse topic structure has wide potential for application in discourse analysis and related tasks, the research on constructing such discourse resources is quite limited in Chinese language. In this paper, we propose a micro-topic scheme (MTS) to represent the discourse topic structure in the Chinese language according to theme-rheme theory, with elementary discourse topic unit(EDTU) as the node and referent of theme-rheme as link. In particular, thematic progression is employed to directly represent the development of the discourse topic structure.

Results: Guided by the MTS, we manually annotate a Chinese Discourse Topic Corpus (CDTC) of 500 documents. Moreover, we get 89.9 and 72.15 F1 value in two identification preliminary experiments, respectively, which show that the proposed representation can perform good automatic computation.

Conclusion: The lack of the formal representation system and related corpus resources for Chinese discourse topic structure has greatly restricted the study of discourse topic analysis in natural language, and further affected the development of natural language understanding. To address the above issues, a micro-topic scheme(MTS) representation is proposed based on functional grammar theory, and the corresponding corpus resources(i.e., CDTC) are constructed. Our preliminary evaluation justifies the appropriateness of the MTS for Chinese discourse analysis and the usefulness of our CDTC.

Keywords: Information extraction, Discourse topic, Discourse analysis, Theme-rheme theory, Thematic progression

Background

It is one of the most challenging tasks for the development of artificial intelligence to make it possible for the machine to understand the text of natural language and even understand the intention of the author. Discourse topic structure analysis is the core work of this task, the main research contents are the analysis of discourse topic structure and semantic relations between the units from the whole text level, and use the context of discourse comprehension.

Discourse topic structure is also the key to the cohesion of the discourse and reflects the essence of the text [1]. Over the last few years, discourse topic structure has been widely studied and proven to be a critical cohesive element at the text level [2–7]. A linear segmentation of texts into proper topic structures may reveal valuable information on, for instance, not only the themes of segments but also the overall thematic structure of the text, and it can subsequently be applied to various text analysis tasks, such as text summarization, information retrieval and discourse analysis [8–10].

Although the discourse topic structure has wide potential for application in discourse analysis and related tasks, the research on constructing such discourse resources is quite limited [2, 3], and the focus has mostly rested on the English language except some other research [11, 12]. However, as far as discourse information structure is concerned, English is typologically different from Chinese: the former is a subject-prominent language, where the subject is an indispensable element in determining sentence patterns, and the latter is a topic-prominent language, where the topic makes an important contribution to generate a sentence [13]. This largely differentiates the discourse topic structures in English and Chinese. Unfortunately, previous studies on discourse topic structure fail to fully reflect this difference.

In order to explore the appropriate Chinese discourse topic structure representation, we proposed a micro-topic scheme (MTS) to represent discourse topic structure in the Chinese language according to theme-rheme theory. Subsequently, an automatic analysis system of MTS was constructed for exploring the automatic recognition of Chinese discourse topic.

To the best of our knowledge, this is the first exploration of the use of theme/rheme as a basic unit of discourse structure analysis and the use of thematic progression as a link of discourse relationship analysis in Chinese discourse. Firstly, this model provides a new way of big data processing, which implements a transformation that converts unstructured data to structured data in text. Furthermore, compared with traditional methods, our model has better computability. Automatic recognition for theme/rheme task is associated with most pop research topics in the area of natural language processing, e.g., POS tagging, semantic role labeling (SRL). Effective research on these tasks contributes to improve the computational performance of our current task more easily.

The rest of this paper is organized as follows. "Related work" section briefly overviews the related work. In "Model" section, we present the MTS according to theme-rheme theory, and describe the construction of the CDTC corpus. In "Methods" section, an automatic analysis method of MTS is proposed. "Results and discussion" section provides the experimental result on the identification of entities of MTS, the crucial step for automatic discourse topic analysis. Finally, "Conclusion" section concludes our work.

Related work

The rhetorical structure and the topic structure are not only interdependent but also complemental in discourse analysis.

For the discourse rhetorical structure, with Rhetorical Structure Theory Discourse Treebank (RST-DT) [2] and Penn Discourse Treebank (PDTB) [3] being the most prevalent over the past decade, the emergence of several English corpus provides resources for the analysis of English discourse. In contrast, there are only a few studies on Chinese discourse annotation [14–17], with a focus on using the existing RST (Rhetorical Structure

Theory) or PDTB frameworks. Recently, Li et al. (2014) proposed a Connective-driven Dependency Tree (CDT) structure as a representation scheme for Chinese discourse structure [18]. With both the advantages of PDTB and RST, CDT meets well the special characteristics of Chinese discourse.

For the discourse topic structure, some studies have begun to focus on the topic level in Chinese discourse topic annotation. The OntoNotes corpus [4] was built on two types of infrastructure, the syntax structure and the predicate-argument structure, which were derived from the Penn Treebank corpus and the Penn PropBank corpus, respectively. In addition, the generalized topic framework [5] defines punctuation clauses as the basic unit of Chinese discourse, and the concepts of the generalized topic and the topic clause is proposed to explicitly describe the topic structure in Chinese discourse. Although both the OntoNotes corpus and the generalized topic framework take into account the special characteristics of Chinese discourse, some issues still remain. For example, there is no suitable representation unit to match different levels of topics. In addition, the lack of sufficient corpus resources to meet the research of Chinese discourse topic analysis is also a serious problem.

Model

Micro-Topic Scheme

In order to explore the discourse relationship, we propose a micro-topic scheme (MTS) to represent the discourse cohesion according to the theme-rheme structure based on functional grammar theory [19], which can be formalized as a triple as below:

$$MTS = (S_n, S_{n+1}, \delta_n)$$

Where $S_n \in T \cup R$, $S_{n+1} \in T \cup R$, T represent the set of themes and R is the set of rhemes in the whole discourse,called Static Entities of MTS by us. $\delta_n \in L$, L is a set of cohesion dynamic relationships of MTS between EDTUs, called Micro-Topic Link(MTL) by us. The visual representation of the model is shown in part (b) of the Fig. 1 below. Some definitions in the model are as follows.

Definition 1 (Elementary Discourse Topic Unit (EDTU)) *is defined as the basic unit of discourse topic analysis, which is limited to clause.*

Inspired by Rhetorical Structure Theory, an EDTU should contain at least one predicate and express at least one proposition. Moreover, an EDTU should be related to other EDTUs with some propositional function. Finally, an EDTU should be punctuated. For Example 1, (a) is a single sentence with serial predicate; (b) is a complex sentence with two EDTUs(clauses).

Example 1

(a) *She started the car. (single sentence, serial predicate, one EDTU)*
(b) *She started the car, and drove off.(complex sentence, two EDTUs)*

In order to improve the computational performance, we give the main structure of Theme and Rheme as defined in Definition 2.

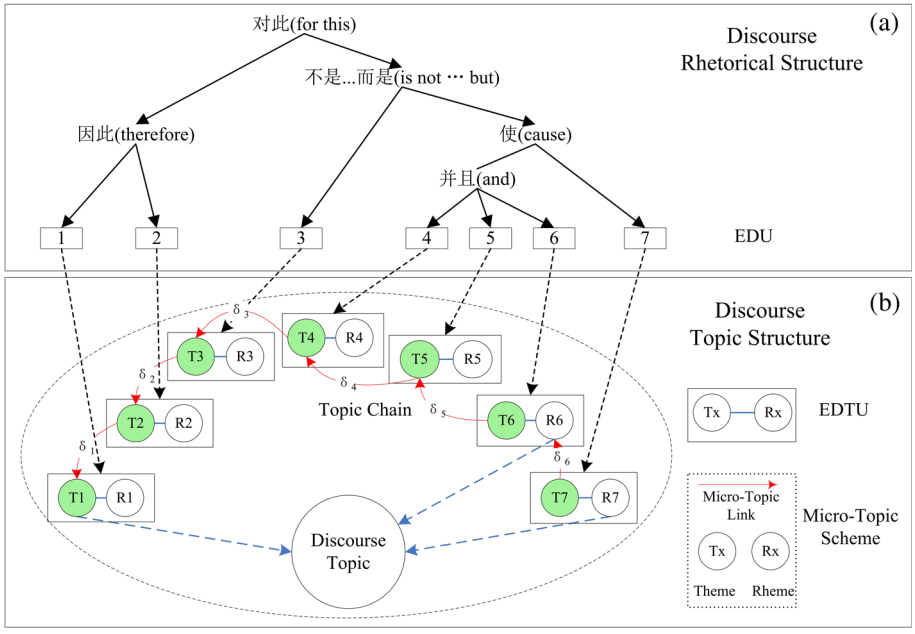

Fig. 1 An example of discourse topic structure in MTS with the corresponding discourse rhetorical structure. According to a connective-driven dependency tree (CDT) scheme, connectives were directly used to represent the hierarchy structure of a CDT and the rhetorical relation of a discourse, as shown in part (**a**) of this Figure. Part (**b**) in this Fig. 1 gives an example of MTS representation, corresponding to Example 10 shown in manuscript. It consists of 7 clauses, excerpted from chtb0001 which is from OntoNotes corpus. Here, a clause is equivalent to an EDTU, which is constituted by a theme and a rheme, denoted by Tx and Rx, respectively

Definition 2 (Theme and Rheme) *Theme Structure is the left part of the predicate in the EDTU for Chinese, and the remainder is Rheme Structure.*

Taking *Example* 1 as an example, we can find that *She* is the Theme, and *started the car* is the Rheme.

Definition 3 (Micro-Topic Link (MTL)) *A MTL is a representation of the semantic association between the themes or rhemes, which are derived from the adjacent EDTUs. This semantic association is expressed as four thematic progression patterns formally, while in content, it reflects the cohesive properties of the discourses, which mainly include* ***reference, ellipsis, substitution, repetition, synonym/antisense, hyponymy, meronymy, and collocation****.*

• **Reference** means that the current theme(or rheme) in an EDTU refers to the previous one.

Example 2

(a) $[这张条子]_{T1}[是(安娜)_{Nucleus}留的]_{R1}$, (b) $[她]_{T2=R1(Nucleus)}$刚才来过。

(a) $[This\ note]_{T1}$ $[was\ left\ by\ [Anna]_{Nucleus}]_{R1}$, (b) $[who]_{T2=R1(Nucleus)}$ *had just come.*

In the above Example 2, EDTU(a) and EDTU(b) constitute a MTS through a MTL, which is a connection of reference between *"安娜 (Anna)"* and *"她 (who)"*. Among them, *"是安娜留的 (was left by Anna)"* is the rheme of EDTU(a), and *"她 (who)"* is the theme of EDTU (b).

• **Ellipsis** means that the theme or rheme of the second EDTU is omitted, which is a kind of grammatical method to avoid repetition, highlight new information, and make the text more compact. As shown in Example 3, theme *"我 (I)"* was omitted in EDTU (b).

Example 3

(a) $[我]_{T1}[早上出门]_{R1}$, (b) $[ZeroA]_{T2=T1}[看到一只猫]_{R2}$.

(a) $[I]_{T1}$ *[went out in the morning]*$_{R1}$ *(b) and* $[ZeroA]_{T2=T1}$*[saw a cat]*$_{R2}$.

• **Substitution** means that the theme(or rheme) in the latter EDTU is replaced by a substitute for words, which has the same meaning as the replaced component. As shown in Example 4, rheme *"新的 (a new one)"* was an substitute word in EDTU (b) for the replaced component, which is *"史蒂夫的帽子 (Steve's hat)"*.

Example 4

(a) $[史蒂夫的帽子]_{T1}[太破了]_{R1}$, (b) $[他]_{T2}[需要换个[新的]_{Nucleus}]_{R2(Nucleus)=T1}$。

(a) $[Steve's\ hat]_{T1}$ *[is too broken]*$_{R1}$. *(b)* $[He]_{T2}$ *[needs [a new one]*$_{Nucleus}]_{R2(Nucleus)=T1}$.

• **Repetition** means that the theme(or rheme) has appeared many times, such as *"熊 (bear)"* in Example 5.

Example 5

(a) $[安吉拉]_{T1}[遇到了一只[熊]_{Nucleus}]_{R1}$, (b) $[熊]_{T2=R1(Nucleus)}$看起来有点饥饿。

(a) $[Algy]_{T1}$ *[met [a bear]*$_{Nucleus}]_{R1}$. *(b)* $[The\ bear]_{T2=R1(Nucleus)}$ *looks a bit hungry.*

• **Synonym/antisense** means that the themes(or rhemes) related to two EDTUs are a pair of synonyms or antonyms. Example 6 shows that *"朋友 (a friend)"* and *"敌人 (enemy)"* is a pair of antonyms.

Example 6

(a) $[朋友]_{T1}[褒扬你的美德]_{R1}$, (b) $[敌人]_{T2=T1}[夸大你的过错]_{R2}$。

(a) $[A\ friend]_{T1}$ *[praises a man's virtue]*$_{R1}$, *[and the enemy]*$_{T2=T1}$ *exaggerates his fault.*

• **Hyponymy** means that the themes(or rhemes) related to two EDTUs form an abstract and concrete relationship. As shown in Example 7, *"狼(wolf)"* is a kind of *"动物(animal)"*.

Example 7

(a) $[狼]_{T1}[一般生活在草原上]_{R1}$, (b) $[这种动物]_{T2=T1}$喜欢群居。

(a) $[The\ wolves]_{T1}$ *[usually live on the grassland]*$_{R1}$, *(b) [and the animals]*$_{T2=T1}$ *like to live in groups.*

• **Meronymy** means that the theme(or rheme) in one EDTU is a part of the theme(or rheme) from the other EDTU. As shown in Example 8, *"他的头发 (his hair)"* is a part of *"一名中年男子 (A middle-aged man)"*, from the point of view of body composition.

Example 8

(a) $[一名中年男子]_{T1}[迎面走来]_{R1}$, (b) $[他的头发]_{T2=T1}$非常光亮。

(a) $[A\ middle\text{-}aged\ man]_{T1}$ *[is walking on the head]*$_{R1}$, *(b)* $[his\ hair]_{T2=T1}$ *is very bright.*

• **Collocation** means that the themes(or rhemes) related to two EDTUs belong to a set of semantically related words. There are two groups of words as follows, for instance, *"ice, snow, white"* and *"night, star"*.

Example 9

(a) [雪]$_{T1}$[下了一夜]$_{R1}$, (b) [整个田野]$_{T2}$[[白]$_{Nucleus}$茫茫一片]$_{R2=T1}$。
(a) [Snow]$_{T1}$ [had fallen all night]$_{R1}$, (b) [while the fields]$_{T2}$ [were a vast expanse of [whiteness]$_{Nucleus}$]$_{R2=T1}$.

In the above Example 9, *"snow"* and *"whiteness"* constitute the MTL, which is a connection between EDTU(a) and EDTU(b).

Definition 4 (Discourse Topic (DT)) *A DT is composed of n MTSs($n \geq 1$), which are connected by MTLs.*

In fact, the DT is a recursive definition, which can be expressed as follows:

Rule (1) A MTS is a **DT**.
Rule (2) Two **DTs** connected with MTL is a **DT**.
Rule (3) **DT** belongs to the union of all sets satisfying Rule (1) and Rule (2).

Definition 5 (Micro-Topic Chain(MTC)) *A MTC is a sequence of connected MTLs, which are contained in a DT.*

The topic chain is a common phenomenon in Chinese. The contextual referring expressions are frequently omitted in Chinese discourse, which leads to the difficulty associated with the topic chain [20]. Typically, in order to enable the reader to find a specific discourse coherence, the referring expression has sufficient topic continuity. And above all, a topic chain will be made up of the identical topics which linked by anaphora (zero anaphora or not) [21].

To illustrate our proposed MTS, we give an Example 10 as below.

Example 10 (1)[[浦东]$_{Satellite}$开发开放]$_{T1}$[是一项振兴上海, 建设现代化经济、贸易、金融中心的跨世纪工程]$_{R1}$, (2) [< ZeroA >$_{Nucleus}$(因此)大量出现的]$_{T2(Nucleus)=T1(Satellite)}$ [是以前不曾遇到过的新情况、新问题]$_{R2}$。(3) [(浦东), 浦东]$_{T3=T2(Nucleus)}$ [不是简单的采取"干一段时间, 等积累了经验以后再制定法规条例" 的做法]$_{R3}$,(4)[< ZeroA >]$_{T4=T3}$[而是借鉴发达国家和深圳等特区的经验教训]$_{R4}$,(5)[< ZeroA >]$_{T5=T4}$ [<并且>聘请国内外有关专家学者]$_{R5}$, (6) [< ZeroA >]$_{T6=T5}$[< 并且 > 积极、及时地制定和推出法规性文件]$_{R6}$, (7)[< ZeroA >]$_{T7=T6}$ [使 这些经济活动一出现就被纳入法制轨道]$_{R7}$。

(1)[Pudong's development and opening]$_{T1}$ [is an undertaking spanning a century for vigorously promoting Shanghai and constructing a modern economic, trade, and financial center]$_{R1}$. (2)Because of this, <during the process of [Pudong's]$_{Satellite}$ development and opening,>$_{ZeroA=T2=T1}$ [new situations and new questions that were not encountered previously are emerging in great numbers]$_{R2}$. (3)[In response to this, Pudong]$_{T3=T2(Satellite)}$ [is not simply adopting an approach of "work for a short time and then draw up laws and

regulations only after experience has been accumulated."]$_{R3}$ (4)[Instead, Pudong]$_{T4=T3}$ [is taking advantage of the lessons from the experience of developed countries and special regions such as Shenzhen]$_{R4}$, (5)[< ZeroA >]$_{T5=T4}$ [by hiring appropriate domestic and foreign specialists and scholars]$_{R5}$, (6)[< ZeroA >]$_{T6=T5}$ [actively and promptly formulating and issuing regulatory documents]$_{R6}$. (7)<According to these documents,>$_{(ZeroA=T7=T6)}$ [these economic activities are incorporated into the sphere of influence of the legal system as soon as they appear]$_{R7}$.

Li [18] proposed a connective-driven dependency tree (CDT) scheme to represent discourse rhetorical structure in the Chinese language, in which elementary discourse units (EDUs) were used as leaf nodes and connectives were used as non-leaf nodes. Especially, connectives were directly used to represent the hierarchy structure of a CDT and the rhetorical relation of a discourse, as shown in part (a) of the Fig. 1.

Part (b) in Fig. 1 gives an example of MTS representation, corresponding to Example 10 shown above. It consists of 7 clauses, excerpted from chtb0001 which is from OntoNotes corpus. Here, a clause is equivalent to an EDTU, which is constituted by a theme and a rheme, denoted by Tx and Rx, respectively. For instance, "*In spite of the fact that of the regulatory documents that the Pudong new region*" stands for the theme in the first clause(a), and the rheme occupies the rest, "*has formulated*".

Similar to what we described above, we define a DT as a set of MTSs sharing an identical topic connected by MTLs. For example, there are two DTs in Example 10, as shown in part (b) of the Fig. 1: $T1 \leftarrow T2 \leftarrow T3 \leftarrow T4 \leftarrow T5 \leftarrow T6$, and $R6 \leftarrow T7$, One MTC is guided by the overt identical NP(Noun Phrase) "浦东$_{Satellite(T1)}$ 开发开放 (T1) (Pudong's development and opening up)", the DT that spreads over six EDTUs (clauses 1∼6). As we can see in Fig. 1, six overt coreferential NPs are considered to form a MTC, with the overt NP (T1) being the head topic of the chain, and the following MTC shares one single topic. In comparison, the other chain refers to the DT "法规性文件 (regulatory documents)" headed by R6 and followed by T7 (zero anaphora).

According to the theme-rheme theory [19], there is a reference relationship between the theme or rheme of current EDTU and previous EDTU. As shown in Part(b) of Fig. 1, an arrow is employed to indicate this reference by pointing to the theme or rheme in the EDTU, such as T2=T1, T3=T2, T4=T3, T5=T4, T6=T5 and T7=R6.

Static Entity of MTS

Derived mainly from the systemic-functional grammar [19], theme and rheme are two static entities representing the way in which information is distributed in a clause. While theme indicates the given information serving as the departure point of a message, which has already been mentioned somewhere in text or shared as mutual knowledge from the immediate context, rheme is the remainder of the message in a clause in which theme is developed.

From the view point of discourse analysis, we are interested in the sequences of thematic and rhematic choices creating certain kinds of thematic patterns instead of the actual individual choices of themes or rhemes. Therefore, our scheme to the notion of theme is discourse-oriented, that is, we are most concerned with the role theme fulfills in constructing and developing a discourse dynamic relationship, as opposed to individual sentences.

Dynamic Relationship of MTS

Previous studies [22–24] have claimed that the way in which lexical strings and reference chains interact with theme/rheme is not random; rather the patterns of interaction realize what they refer to as a text's thematic progression. Figure 2 shows four major dynamic relationships of thematic progression proposed in the literature:

(I) Constant Progression, where the theme of the subsequent clause is semantically equivalent to the theme of the first clause.

Example 11 *(a) Two beggars (T1) had been hiding (R1). (b)They(T2=T1) saw the money (R2).*

(II) Centralized Progression, where the rheme of the subsequent clause is semantically equivalent to the rheme of the first clause.

Example 12 *(a) The children (T1) laughed (R1). (b) Then their mother(T2) laughed, too (R2=R1).*

(III) Simple Linear Progression, where the theme of the subsequent clause is semantically equivalent to the rheme of the first clause.

Example 13 *(a)Our school (T1) is a big garden (R1). (b)In the garden(T2=R1) grow many flowers (R2).*

(IV)Crossed Progression, where the rheme of the subsequent clause is semantically equivalent to the theme of the first clause.

Example 14 *(a) The exhibition (T1) was good (R1). (b) I (T2) liked it very much (R2=T1).*

As shown in Example 10, constant progression is suitable for the referent relationships among clauses 1-6.

Corpus building based on MTS

Based on this MTS model, we annotated a Chinese discourse topic corpus(CDTC) with 500 discourses from OntoNotes corpus English datasets(chtb0001-chtb0325, chtb0400-chtb0657). To begin with, the same dataset were annotated by two annotators simultaneously. Moreover, we calculate the consistency of annotations with Kappa Value by using the two sets of annotated data. Finally, a formal corpus dataset will be constructed and checked by senior tagging instructor. Table 1 illustrates the inter-annotator consistency specifically. The CDTC is also used for our experiment as dataset.

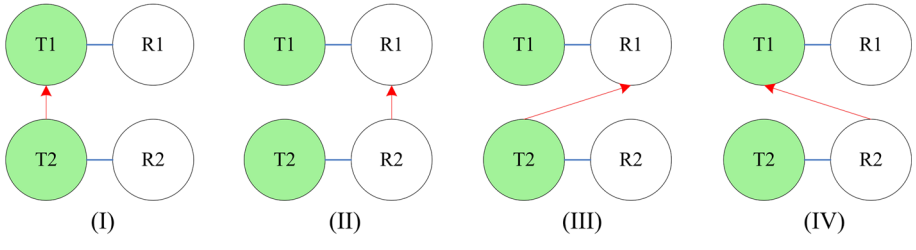

Fig. 2 Four dynamic relationships of thematic progression. This figure shows four major dynamic relationships of thematic progression proposed in the literature, e.g., Constant Progression, Centralized Progression, Simple Linear Progression, Crossed Progression

Table 1 Inter-annotator consistency

	Agreement%	Kappa
EDTU	96.0	0.91
Static entities of MTS	92.0	0.83
MTL	89.0	0.86

Methods

Overall processing pipeline

To evaluate the computability of our CDTC corpus, we present the experimental results on the identification of static entity of MTS(i.e., theme-rheme structures), which is a crucial component of discourse topic analysis.

Our model framework is summarized in Fig. 3. This system takes an input discourse and output the confidence score of the entity of MTS. It primarily consists of the following three components: Inputting the discourse, Identifying the EDTUs (Elementary Discourse Topic Units) and Identifying static entity of MTS. To begin with, the input of the system is the discourse from natural language without any preprocessing. In succession, the comma is used as a boundary sign, and the classifier model is obtained

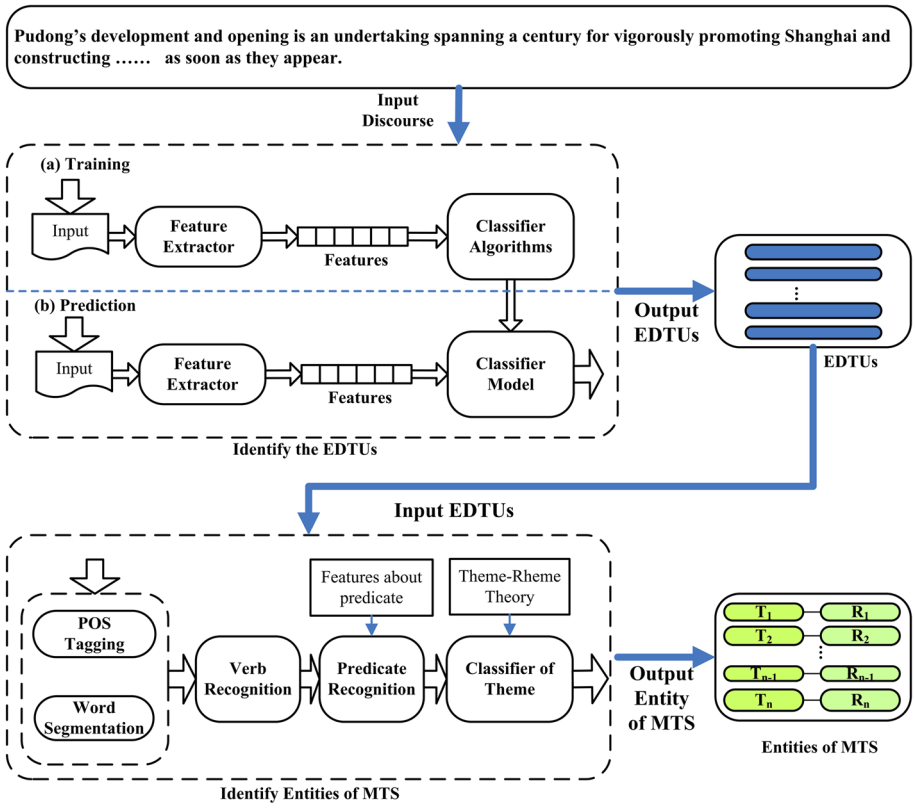

Fig. 3 The Experimental Framework of Identifying entities of MTS. This system according to which takes an input discourse and output the confidence score of the entity of MTS. It primarily consists of the following three components: Inputting the discourse, Identifying the EDTUs(Elementary Discourse Topic Units) and Identifying static entity of MTS. To begin with, the input of the system is the discourse from natural language without any preprocessing. In succession, the input text is split into a number of simple sentences containing a predicate, which is defined as EDTUs. Finally, according to these EDTUs, entities of MTS in which are identified

by machine learning algorithms. With the help of this classifier model, the input text is split into a number of simple sentences containing a predicate, which is defined as EDTUs. Finally, according to these EDTUs, entities of MTS in which are identified. Take Example 10 as an instance, we will describe each components in our model as below.

Identifying the EDTUs

According to the Definition 1, the Example 10 has 7 EDTUs, which include Clause(a), (b), (c), (d), (e), (f) and (g).

For the automatic identification of EDTU, inspired by Li [18], we consider this as a binary classification for EDTU's boundary and use some machine learning methods to solve this problem. We used various features listed in Table 2 specifically, which had adopted in [25] and [18]. Table 3 shows the performance of EDTU identification on the CDTC with 10-fold cross validation via the Mallet toolkit [26].

Identifying entities of MTS

According to Definition 2 of subsection Model, the Example 10 has 7 themes and 7 rhemes, which are represented by T1-T7 and R1-R7, respectively.

For the automatic identification of entities of MTS, according to our Definition 2, the predicate is used as a division sign, and thus, the identification of entities of MTS is equivalent to the predicate identification problem in a way. In other words, this problem is transformed into a classical semantic role labeling problem. It is worth noting that the recognition of the predicate is limited to one EDTU range, which would contribute to a better recognition result. In addition to classical predicate features in previous studies [27–29], more features are derived from nominal and verbal SRL(Sematic Role Labeling), such as the location in terms of the NP, the path features, intervening verb and the arguments. Using the Mallet toolkit [26] with features listed in Tables 4 and 5 shows the performance of identifying the entities of MTS on CDTC corpus with 10-fold cross validation.

Table 2 Features of EDTUs identification

Name	Description
POS_Pre_Word	Part of speech tagging for the previous word
Rep_Pre_Word	A string representation for the previous word
POS_Foll_Word	Part of speech tagging for the following word
Rep_Foll_Word	A string representation for the following word
Left_Phrase_Label	Left brother's phrase label
Right_Phrase_Label	Right brother's phrase label
Con_Phrase_Label	Conjunction of phrase label of left brother and right brother
Con_Family_Label	Conjunction of the ancestors and Con_Phrase_Label
Is_Sub_Conjunction	Is there a subordinating conjunction for left of the comma?
Is_CoordIP	Is the parent of the comma a coordinating IP construction?
Is_Top_Child	Is the comma a top-level child?
Is_Top_CoordIP	Is the parent of the comma with top-level child and coordinating IP construction?
Pun_Mark_Temp	Punctuation mark template of this sentence
Distance_Left_Right	Length difference between the left and right segments of the comma

Table 3 Performance of EDTUs identification

Classifier	Gold			Automatic		
	Precision	Recall	F1	Precision	Recall	F1
C45	90.6	90.9	90.5	89.3	90.3	88.6
Naive Bayes	90.3	89.6	89.4	88.5	89.2	87.8
MaxEnt	91.2	90.3	91.9	90.2	90.7	89.9

Results and discussion

Result

Tables 3 and 5 show the result of automatic recognition for the EDTUs and the entities of MTS, respectively.

On the one hand, in order to reflect the independent performance of each module, we extract features from the previous module's manual tagging as input to the current module. This is called Gold shown in Tables 3 and 5. On the other hand, in order to observe the performance of the overall system, we also use the features automatically acquired by the previous module as input to the current module. This is called Automatic.

As shown in Table 3, we obtained some high F1 values based on the Gold data set, the highest one among which reached 91.9%. Meanwhile, the results based on Automatic data set are also very close to the Gold's ones, the highest one among which reached 89.9%. The main reason may lie in the clarity of definition of EDTU and less ambiguity.

As shown in Table 5, compared with the model for recognition of the EDTUs, the performance of the module for recognition of the entities of MTS has decreased. This is not only due to the introduction of errors from the previous module, but also owing to the complexity of identifying the entities of MTS. Despite that, MaxEnt performs the best,

Table 4 Features of MTS's entities identification

Name	Description
Predicate	A content word (lemma) of the predicate of each clause
Predicate class	The verb class that the predicate belongs to
Head word	String representation of head word of one clause
POS of head word	Part of speech of head word
Phrase type	Syntactic category of the constituent
Path of span	The path from the span to the nominal predicate
Position	The positional relationship of the span with the predicate, "left" or "right"
Focus word	First word and last word of the focus span
Focus span space	Is the focus span adjacent to the predicate? Yes or No
IsBrothers	Has the predicate brothers? Yes or No
IsRightBrother	Has the predicate right brother? Yes or No
Head word of right brother	The headword of the predicate's nearest right brother
POS of right brother	The POS of the predicate's nearest right brother
IVerb	Intervening verb itself
IVerb class	The verb class that contains IVerb
Path of IVerb	The path from the IVerb to the focus constituent
IsFocusSpArg	Is the focus span an argument for IVerb? Yes or No
Sematic role of focus	The sematic role of the focus span for IVerb
IsHNPArg	Is HNP(Hightest NP headed by the nominal predicate) an argument for IVerb? Yes or No
Sematic role of HNP	The semantic role of HNP for IVerb

Table 5 Performance of MTS's entities identification

Classifier	Gold			Automatic		
	Precision	Recall	F1	Precision	Recall	F1
C45	76.5	77.4	76.95	68.3	66.5	67.39
Naive Bayes	76.1	76.9	68.8	67.9	78.2	68.35
MaxEnt	79.8	80.3	80.05	72.5	71.8	72.15

with a F1 measure as high as 80.05% on gold data and a F1 measure as high as 72.15% on automatic data.

In Summary, the result suggests the appropriateness of our definition of the micro-topic scheme.

Discussion

The importance of MTS lies in constructing a suitable representation for computing the discourse topic. The specific analysis is as follows:

(a) The unified definition of EDTU is consistent with EDU from Rhetorical Structure Theory (RST), which provides the basis for discourse analysis through the joint research of discourse topic structure and discourse rhetorical structure.

(b) The formal definition of MTL involves incorporating a variety of cohesive relations into the scope of semantic relations, which provides a more complete research content for the study of the discourse semantic relations.

(c) The recursive definition of the discourse topic (DT) reflects the level of the topic, which provides a basis for the hierarchical research of discourse topic structure.

(d) In the implementation of MTL, the patterns of thematic regression are introduced, which provide a dynamic evolution process for text generation. In other words, it provides a computable model for text generation.

In sum, (d) is a dynamic analysis process, and (a), (b) and (c) achieve a static representation architecture. On the basis of the combination of the above, the MTS provides a full representation system and a suitable deductive tool for discourse analysis.

Conclusion

In this paper, we propose a micro-topic scheme (MTS) as a representation for Chinese discourse topic structure according to theme-rheme theory. MTS has the advantages of both the OntoNotes corpus and the generalized topic framework and adapts well to the special characteristics of Chinese discourse. Especially, we analyzed the characteristics of MTS in a comprehensive way from the various perspectives of EDTU, Static Entity of MTS(i.e.,theme-rheme structure), Dynamic Relationship of MTS(i.e.,micro-topic link) and micro-topic chain. Based on the MTS scheme, we annotate 500 documents according to a top-down segmentation and chain-backtracking strategy to remain consistent with a Chinese native's cognitive habits. Evaluation of the CDTC corpus proves the appropriateness of the MTS scheme for Chinese discourse cohesion structure and the usefulness of our CDTC corpus.

Abbreviations

CDT: Connective-driven dependency tree; CDTC: Chinese discourse topic corpus; DT: Discourse topic; EDU: Elementary discourse unit; EDTU: Elementary discourse topic unit; MaxEnt: maximum entropy model MTS: Micro-topic scheme; MTL: Micro-topic link; MTC: Micro-topic chain; NP: Noun phrase; PDTB: Penn discourse treebank; RST: Rhetorical structure theory; RST-DT: rhetorical structure theory discourse Treebank; SRL: Sematic role labeling

Acknowledgements
This work was supported by the National Natural Science Foundation of China (No.61331011, No.61673290) and Foundation of Key Laboratory in Science and Technology Development Project of Suzhou (No. SZS201609).

Funding
Not applicable.

Availability of data and materials
There is a major part of patent protection of software, and therefore cannot be available online. Some data sets will be public in the near future to allow for repeated results.

Authors' contributions
X-fX is the co-designer and software developer of the system, GZ is the co-designer of the system, and the academic advisor for X-fX, two authors have contributed to the write-up of this paper. Both authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Department of Computer Science and Technology, Suzhou University of Science and Technology, KeRui Road, Suzhou, China. [2]School of Computer Science and Technology, Soochow University, ShiZi Road, Suzhou, China. [3]Virtual Reality Key Laboratory of Intelligent Interaction and Application Technology of Suzhou, KeRui Road, Suzhou, China.

Received: 26 May 2017 Accepted: 6 August 2017
Published online: 07 September 2017

References

1. De Beaugrande RA, Dressler WU, Vol. 1. Introduction to Text Linguistics. London: Longman; 1981.
2. Carlson L, Marcu D, Okurowski ME. Building a discourse-tagged corpus in the framework of rhetorical structure theory In: van Kuppevelt J, Smith RW, editors. Current and New Directions in Discourse and Dialogue. Dordrecht: Springer; 2003. p. 85–112.
3. Prasad R, Dinesh N, Lee A, Miltsakaki E, Robaldo L, Joshi AK, Webber BL. The penn discourse treebank 2.0. In: Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech: DBLP; 2008. p. 2961–8.
4. Weischedel R, Pradhan S, Ramshaw L, Kaufman J, Franchini M, El-Bachouti M, Xue N, Palmer M, Marcus M, Taylor A, Greenberg C, Hovy E, Belvin R, Houston A. OntoNotes Release 4.0. Philadelphia: Linguistic Data Consortium; 2010.
5. Song R, Jiang Y, Wang J. On generalized-topic-based Chinese discourse structure. In: CIPS-SIGHAN Joint Conference on Chinese Language Processing. Stroudsburg: ACL Press; 2010. p. 23–33.
6. Zhou G, Li P. Improving syntactic parsing of Chinese with empty element recovery. J Comput Sci Techn. 2013;28(6): 1106–1116.
7. Rutherford A, Xue N. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In: Proceedings of the 14th Annual Conference of the North American Chapter of the ACL-HLT. Stroudsburg: ACL Press; 2015. p. 799–808.
8. Salton G, Singhal A, Buckley C, Mitra M. Automatic text decomposition using text segments and text themes. In: Proceedings of the Seventh ACM Conference on Hypertext. Washington: ACM Press; 1996. p. 53–65.
9. Du L, Buntine WL, Johnson M. Topic segmentation with a structured topic model. In: Proceedings of the 12th Annual Conference of the North American Chapter of the ACL-HLT. Stroudsburg: ACL Press; 2013. p. 190–200.
10. Galley M, McKeown K, Fosler-Lussier E, Jing H. Discourse segmentation of multi-party conversation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL Press; 2003. p. 562–9.
11. Ren CY. A Grammar of Spoken Chinese. Berkeley and LosAngeles: University of California Press; 1968.
12. ChengXi Q. Chinese Discourse Grammar. Beijing(in Chinese): Beijing language and culture university press; 1998.
13. Li CN, Thompson SA. Subject and topic: A new typology of language. New York: Academic Press; 1976. pp. 457–89.
14. Chen L. English and Chinese discourse structure dimension theory and practice. Shanghai: PhD thesis, Shanghai International Studies University; 2006.
15. Ming Y. Rhetorical structure annotation of Chinese news commentaries. J Chin Inf Process. 2008;4:2–11.
16. Xue N. Annotating discourse connectives in the Chinese treebank. In: Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky. Washington: ACM Press; 2005. p. 84–91.
17. Zhou Y, Xue N. Pdtb-style discourse annotation of Chinese text. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press; 2012. p. 69–77.

18. Li Y, Feng W, Sun J, Kong F, Zhou G. Building Chinese discourse corpus with connective-driven dependency tree structure. In: EMNLP. Stroudsburg: ACL Press; 2014. p. 2105–114.
19. Halliday MAK, Matthiessen CMIM. An Introduction to Functional Grammar. London: Hodder Education; 2004.
20. Yeh CL, Chen YC. 442 Zero anaphora resolution in Chinese with partial parsing based on centering theory. In: Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering. Piscataway: IEEE Press; 2003. p. 683–8.
21. Li W, Vol. 57. Topic Chains in Chinese: A Discourse Analysis and Applications in Language Teaching. München: Lincom Europa; 2005.
22. Dañes F. Functional sentence perspective and the organisation of text In: Dañes F, editor. Papers on Functional Sentence Perspective. The Hague: Mouton; 1974. p. 106–28.
23. Fries PH. On the status of theme in English: Arguments from discourse In: Petőfi JS, Sözer E, editors. Micro and Macro Connexity of Texts. Hamburg: H. Buske; 1983. p. 116–52.
24. Zhu Y. Patterns of thematic progression and text analysis. Foreign Lang Teach Res. 1995;3:6–12.
25. Xue N, Yang Y. Chinese sentence segmentation as comma classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press; 2011. p. 631–5.
26. McCallum AK. Mallet: A machine learning for language toolkit. 2002. http://mallet.cs.umass.edu.
27. Jiang ZP, Ng HT. Semantic role labeling of nombank: A maximum entropy approach. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press; 2006. p. 138–45.
28. Li J, Zhou G, Zhao H, Zhu Q, Qian P. Improving nominal srl in Chinese language with verbal srl information and automatic predicate recognition. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press; 2009. p. 1280–8.
29. Yang H, Zong C. Multi-predicate semantic role labeling. In: Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press; 2014. p. 363–73.